Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: https://orca.cardiff.ac.uk/id/eprint/99163/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Zhang, Wei and Liu, Hantao ORCID: https://orcid.org/0000-0003-4544-3481 2017. Towards a reliable collection of eye-tracking data for image quality research: challenges, solutions and applications. IEEE Transactions on Image Processing 26 (5) , pp. 2424-2437. 10.1109/TIP.2017.2681424 file

Publishers page: http://dx.doi.org/10.1109/TIP.2017.2681424 <http://dx.doi.org/10.1109/TIP.2017.2681424>

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

information services
gwasanaethau gwybodaeth

Toward a Reliable Collection of Eye-Tracking Data for Image Quality Research: Challenges, Solutions, and Applications

Wei Zhang, *Student Member, IEEE*, and Hantao Liu, *Member, IEEE*

***Abstract*— Image quality assessment potentially benefits from the addition of visual attention. However, incorporating aspects of visual attention in image quality models by means of a perceptually optimized strategy is largely unexplored. Fundamental challenges, such as how visual attention is affected by the concurrence of visual signals and their distortions; whether visual attention affected by distortion or that driven by the original scene only should be included in an image quality model; and how to select visual attention models for the image quality application context, remain. To shed light on the above unsolved issues, designing and performing eye-tracking experiments are essential. Collecting eye-tracking data for the purpose of image quality study is so far confronted with a bias due to the involvement of stimulus repetition. In this paper, we propose a new experimental methodology to eliminate such inherent bias. This allows obtaining reliable eye-tracking data with a large degree of stimulus variability. In fact, we first conducted 5760 eye movement trials that included 160 human observers freely viewing 288 images of varying quality. We then made use of the resulting eye-tracking data to provide insights into the optimal use of visual attention in image quality research. The new eye-tracking data are made publicly available to the research community.**

***Index Terms*— Visual attention, fixation, eye-tracking, image quality, saliency, gaze.**

I. Introduction

DIGITAL imaging systems generate, as a side effect, various types of distortion in visual signals [1]. Visual distortions degrade the quality of digital media content and consequently, may affect consumers' visual experiences or lead to analytical errors in visual inspection tasks [2], [3]. To prevent the appearance of visual distortions and to control image quality, current imaging systems rely on algorithms that can automatically predict image quality as perceived by human observers. The basis of these algorithms is formed by the so-called objective quality metric (OQM).

Substantial progress has been made on the development of OQMs. The state of the art OQMs mainly benefit from the advances in understanding and modelling early visual processing in the human visual system (HVS) and its underlying quality perception behaviour [4]–[6]. Significant findings in visual psychophysics, such as contrast sensitivity and masking have been mathematically modelled and integrated in various OQMs [7]–[12]. By incorporating functional aspects of the HVS, distortion can be quantified in a way that reflects its genuine annoyance to the human eye, which consequently results in a more reliable image quality prediction.

A significant trend in current image quality research is to investigate the impact of visual attention, which is an essential aspect of the HVS. Visual attention refers to a mechanism that enables the HVS to select the most relevant information in a visual scene [13]. Such attentional selection is known to be guided by two types of mechanism, namely the stimulus-driven, bottom-up mechanism and the expectation-driven, top-down mechanism [13]. In the area of computer vision, visual attention is mainly concerned with the former attentional mechanism, and is often interchangeably referred to as saliency [14]–[24]. The empirical foundation of saliency modelling lies in the eye movements of human observers [25]–[28]. Computational models of visual saliency (i.e., bottom-up attentional mechanism) aim at explicitly addressing the first few seconds of eye movements in free-viewing a visual stimulus [13]. A saliency model generally outputs a topographic map that represents conspicuousness of scene locations, where some parts of a scene that appear to an observer to stand out relative to their neighbouring parts.

Incorporating saliency has demonstrated great potential for further improvement of OQMs [29]–[31]; however, finding ways to achieve such integration in a perceptually optimised way remains largely unexplored. The challenge lies in the fact that our knowledge about how saliency is actually affected by the concurrence of visual signals and their distortions as well as the associated implications for image quality judgements is very limited. Due to the lack of such knowledge, the vast majority of existing work has focused on simply utilising a specific saliency model as a weighting function to improve a specific OQM [32]–[36]. However, the following issues such as how to optimise the combination of saliency and OQMs and how to determine appropriate saliency models remain, which are the urgent topics to be investigated.

II. Related Work and Contributions

A. Related Work

Psychophysical studies have been attempted to better understand visual saliency in relation to image quality assessment [37]–[43]. For example, an eye-tracking study was performed in [40] to investigate (via visual inspection of fixation patterns) how task-free fixations (i.e., saliency)

Manuscript received May 18, 2016; revised January 9, 2017; accepted February 22, 2017. Date of publication March 9, 2017; date of current version April 1, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Kalpana Seshadrinathan.

The authors are with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K. (e-mail: zhangw71@cardiff.ac.uk; hantao.liu@cs.cardiff.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2681424

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/

of undistorted images may be affected by two variables, i.e., quality rating task and visual distortion. Based on the visualisations of eye-tracking data, white noise and blurring (under quality rating conditions) are not observed to significantly impact the fixation patterns (relative to the task-free conditions), whereas the impact tends to be more obvious in the case of compression artifacts. In [41], task-free eye-tracking experiments were conducted to investigate how JPEG compression affects fixations. It shows that the impact of JPEG artifacts on fixations is more disruptive at low image quality than the high quality. The eye-tracking data in [42] indicate that fixations change as visual distortion occurs, and that the extent of the change seems to be more related to the strength of artifacts rather than the type of artifacts. In general, psychophysical studies reveal that visual distortions may lead to a deviation from the natural scene saliency, and that such deviation tends to depend on the visual content, the type of distortion and the level of distortion.

Notwithstanding the above effort, it should be noted that the generalisability of the findings reported in these studies remains limited by the choices made in their experimental design. For example, some experiments used a limited number of human subjects [38]; some experiments were restricted to a small degree of stimulus variability in terms of scene content, distortion type and degradation level [40]–[42]; and some eye-tracking studies involved top-down aspects of visual attention (e.g., the involvement of a quality rating task) rather than studying free-viewing bottom-up saliency [42], [43].

Apart from the above drawbacks, existing studies by their nature potentially suffer from an inherent bias due to the involvement of stimulus repetition. Typical eye-tracking data collection for the purpose of image quality assessment often involves each observer viewing the same scene repeatedly several times (with multiple variations of distortion) throughout a session. This repetition (i.e., repeated versions of the same scene) becomes massive as the number of distortion types and/or levels increases and would potentially skew the intended eye-tracking data. In [44], eye-tracking data were collected where participants first viewed 12 short videos and then after a 2-min break they viewed the same 12 videos again. The results showed that there was a notable difference in the locations of the participants' gaze for the first and second viewings of the same video. The eye-tracking experiments in [45] included 10 original videos and their 50 impaired versions (i.e., five levels of degradation per original). The results showed evidence for a memory or learning effect for several viewings of the same video content, and that the observers' gaze behaviour tended to be affected by the involvement of stimulus repetition. Both studies suggest that to ensure the consistency of oculomotor behaviour throughout the experiment (i.e., observing stimuli naturally rather than being forced to learn where to look for visual artifacts, e.g.) and as such to guarantee the reliability of fixation data collection, there is a need for reducing the impact of stimulus repetition.

B. Contributions of the Paper

1) Recent literature [46], [47] has revealed potential limitations of existing approaches taken to integrate saliency to OQMs, and the need to investigate the real interactions between natural scene saliency and visual distortions via eye-tracking. To ensure the validity of fixation data collection, we propose a new experimental methodology with carefully justified control mechanisms. This methodology allows reliably obtaining a substantial eye-tracking data with a large degree of stimulus variability in terms of scene content, distortion type as well as degradation level.

2) Unlike previous eye-tracking studies that have focused more on a limited dataset and rather qualitative analysis, the resulting eye-tracking data enable us to thoroughly evaluate the relation between saliency and distortion. In particular, we perform an exhaustive statistical analysis to provide a comprehensive view of the extent to which different types of distortion with each represented at different levels of degradation can actually affect fixation deployment.

3) Up until now, little has been known about how to optimise the integration of saliency and OQMs in a perceptually meaningful way. An important question has arisen whether saliency derived from an original natural scene or that from the same scene affected by unnatural artifacts should be included in OQMs. Based on our eye-tracking data, we assess whether the difference between both types of saliency is sufficiently large to actually affect the performance gain for existing OQMs.

4) Being able to effectively apply saliency in OQMs requires pre-screening of saliency models, since the effectiveness of saliency models differs in different application domains. On the basis of our eye-tracking data, we benchmark the state of the art saliency models for the purpose of image quality assessment. We explicitly evaluate whether these saliency models possess sufficient capabilities of detecting natural scene saliency and its deviation due to quality changes, and the added value of modelled saliency to OQMs.

5) Moreover, we have made the eye-tracking data publicly available [48] to facilitate research on saliency modelling in image quality assessment.

III. Eye-Tracking: Refined Experimental Methodology

Unlike previous studies, our experiment contains a large degree of stimulus variability in terms of scene content, distortion type as well as distortion level. In addition, a dedicated protocol is devised to eliminate potential bias due to the involvement of massive stimulus repetition, which inherently occurs in a typical image quality study. An eye-tracking database was collected with 160 human observers and 288 test stimuli, and from 5760 eye movement trials.

A. Stimuli

A set of test stimuli is constructed by systematically selecting images from a widely recognised image quality assessment database (i.e., LIVE database [49]).

1) Construction of Source Images: from the fixation deployment perspective, natural scenes can be classified based on the degree of saliency dispersion [31]. As the observation revealed from eye-tracking studies in [50] and [51], if an image contains highly salient objects, then most viewers will concentrate their

Fig. 1. Illustration of source images with different degrees of saliency dispersion used in our experiment, which yield 288 test images.

fixations around them, whereas if there is no obvious object-of-interest viewers' fixations will appear as a more evenly distributed pattern. Thus, images with salient objects tend to have less variation in fixations between viewers than images without salient objects. By use of eye-tracking data in [31], the degree of saliency dispersion—the degree of agreement between observers for human fixations—was determined and used to categorise all source images in the LIVE database. The results showed that the majority of images (i.e., 19 out of 29) clustered around the range of medium degree of saliency dispersion. To mitigate the unbalanced distribution of source images, we decided to remove some images having a medium degree of saliency dispersion. This yielded a rather balanced set of 18 source images as illustrated in Fig. 1. The new make-up consists of 6 images of a small degree of saliency dispersion (e.g., images with distinct foreground/background configurations); 4 images of a greater saliency dispersion (e.g., images without any specific object-of-interest); and 8 images that fall into the range of medium degree of saliency dispersion.

2) Construction of Test Images: Test stimuli used in our experiment cover the full range of distortion types available in the LIVE database, including white noise (WN), JPEG compression (JPEG), Gaussian blur (GBLUR), JPEG2000 compression (JP2K) and simulated fast-fading in wireless channels (FF). For each distortion type, three distorted versions per source image were systematically selected, which were intended to reflect three distinct levels of perceived quality: "High" (i.e., with perceptible but not annoying artifacts), "Medium" (i.e., with noticeable and annoying artifacts) and "Low" (i.e., with very annoying artifacts). Taking advantage of the LIVE database that contains per image a "ground truth" mean opinion score (i.e., DMOS), distortion strengths/levels were adjusted perceptually by using the following mapping: DMOS = [10, 40] to "High" quality, DMOS = [40, 70] to "Medium" quality and DMOS = [70, 100] to "Low" quality. By doing so, for a specific distortion type, the selected 18 "High" quality versions of source images are meant to have approximately the same perceived quality; and similarly for other distortion levels (i.e., "Medium" and "Low"). In addition, a "High" quality version of any source image chosen under a specific distortion type is meant to have approximately the same perceived quality as the "High" quality version of the same source image chosen under any other distortion type; and similarly for other distortion levels (i.e., "Medium" and "Low"). The selection procedure resulted in a set of **288** test stimuli (including the originals) from the LIVE database. Fig. 2 illustrates the average DMOS of images (i.e., 90 images

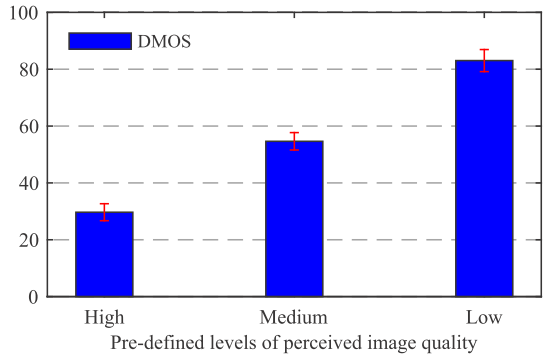


Fig. 2. Illustration of average DMOS of images assigned to a pre-defined level of distortion. The distortion levels are meant to reflect three perceptually distinguishable levels of image quality (i.e., denoted as "High", "Medium" and "Low"). The error bars indicate a 95% confidence interval.

based on 18 source images × 5 distortion types) assigned to individual distortion levels. It clearly shows three distinct means of DMOS (i.e., 30, 55 and 83 within the score range [0, 100]); and hypothesis testing (i.e., based on t-test preceded by a test for the assumption of normality) reveals that the difference between these three pre-defined categories is statistically significant (i.e., with $P < 0.01$ at the 95% confidence level).

B. Proposed Experimental Protocol

There is little consensus on which method is the most appropriate for the conduct of an eye-tracking experiment for the purpose of image quality study. A within-subjects method, in which the same group of subjects views all test stimuli, is commonly used in relevant studies [29], [40]–[42]. This experimental methodology, however, potentially contaminates the results due to carry-over effects, which refer to any effect that carries over from one experimental condition to another [52]. Such effects become more pronounced as the number of test stimuli and/or the rate of stimulus repetition increase in eye-tracking. In our case, the test dataset contains a total of 288 stimuli representing 16 repeated versions (i.e., 15 distorted + 1 original) per source image, which makes the use of a within-subjects method prone to undesirable effects such as fatigue, boredom and learning from practice and experience, and thus increases the chances of skewing the results. To overcome these problems, an alternative method, namely between-subjects [53] was employed in our experiment. In a between-subjects method, multiple groups of subjects are randomly assigned to partitions of test stimuli,

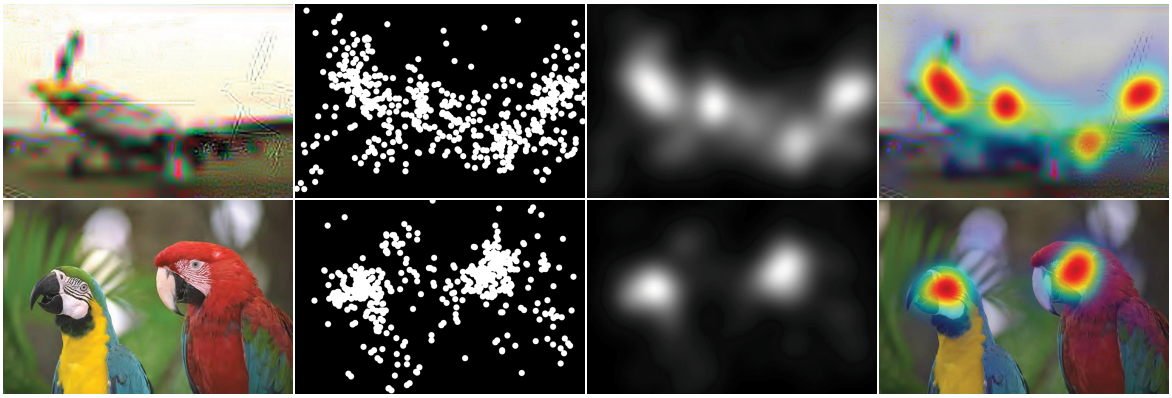


Fig. 3. (a) Two sample stimuli of distinct perceived quality (DMOS = 95.96 (top image) and DMOS = 32.26 (bottom image)). (b) The collection of human eye fixations over 20 subjects. (c) Gaze maps (the darker the regions are, the lower the saliency is). (d) Saliency superimposed on the sample stimuli.

each contains little or no stimulus repetition. We decided to divide the test dataset into 8 partitions of 36 stimuli each; and to allow only 2 repeated versions of the same scene in each partition. To further reduce the carry-over effects, each session per subject was divided into two sub-sessions with a "washout" period between sub-sessions; and by doing so, each subject effectively had to view 18 stimuli without no stimulus repetition in a separate session. Mechanisms were further applied to control the order in which participants per group perform their tasks: (1) half of the participants view the first half partition of stimuli first, and half of the participants view the second half partition first; (2) the stimuli in each sub-session are presented to each subject in a random order. A dedicated control mechanism was also adopted in each sub-session to deliberately include a mixture of all distortion types and the full range of distortion levels. We recruited 160 participants in our experiment, consisting of 80 male and 80 female university students and staff members (between 19 to 42 years of age), all inexperienced with image quality assessment and eye-tracking. The participants were not tested for vision defects, and we considered their verbal expression of the soundness of vision was adequate. The participants were first randomly divided into 8 groups of equal size, each with 10 males and 10 females; and the 8 groups of subjects were then randomly assigned to 8 partitions of stimuli. Based on the rule of thumb for determining sample size in relevant studies (i.e., 5-15 subjects per test stimulus), we assume 20 per stimulus is an adequate sample size (note that the validity of sample size will be further quantitatively tested in Sec. IV).

C. Experimental Procedure

We set up a standard office environment as to the recommendations of [54] for the conduct of our experiment. The test stimuli were displayed on a 19-inch LCD monitor (native resolution is 1024×768 pixels). The viewing distance was set to be approximately 60cm. Eye movements were recorded using an image processing based contact-free tracking system with sufficient head movement compensation (SensoMotoric Instrument (SMI) RED-m). The eye tracking system features a sampling rate of 120Hz, a spatial resolution of 0.1 degree and a gaze position accuracy of 0.5 degree. Each subject was provided with instructions on the purpose and general procedure of the experiment before the start of the actual experiment. Each session per subject contained two successive sub-sessions with a break of 60 minutes between sub-sessions. Since each subject had only two viewings of the same scene, the 60-minute "washout" period was considered sufficient to balance between further reducing the carry-over effects and completing the entire data collection within a reasonable timescale. Each individual sub-session was preceded by a 9-points calibration of the eye-tracking equipment. The participants were instructed to look at the stimuli in a natural way ("view it as you normally would"). Each stimulus was shown for 10 seconds followed by a mid-gray screen of 3 seconds.

IV. Experimental Results

A. Gaze Map

A gaze map representative for stimulus-driven, bottom-up visual attention is derived from the recorded fixations [29]–[31]. Fixations were extracted from the raw eye-tracking data using the SMI BeGaze Analysis Software with minimum fixation duration threshold set to 100ms. A fixation was rigorously defined by SMI's Software using the dispersal and duration based algorithm established in [55]. Fig. 3(b) illustrates the collection of fixations over all subjects (i.e., 20) for each of the two sample stimuli. To construct a topographic gaze map for an average human observer, each fixation location (contained in the aggregated data as shown in Fig. 3(b)) gives rise to a gray-scale patch that simulates the foveal vision of the HVS. The activity of the patch is modelled as a Gaussian distribution of which the width approximates the size of the fovea (2 degree of visual angle). As treated similarly in relevant literature (see e.g., [29], [41], [42]) the duration of fixation was not included when creating a gaze map.

B. Validation: Proposed Reliability Testing

Since standardised methodology for the collection of eye-tracking data does not exist, researchers often follow best practice guidelines for the design of their own experiments. The resulting data, however, differ in their reliability

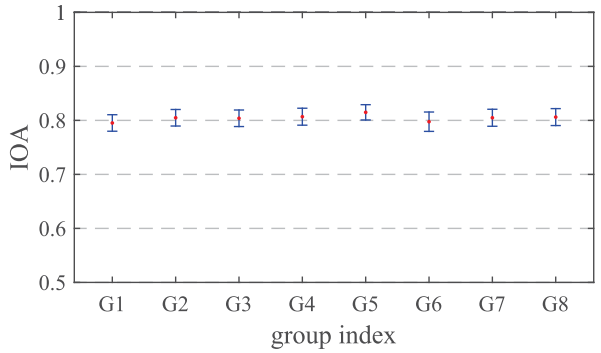


Fig. 4. Illustration of inter-observer agreement (IOA) value averaged over all stimuli assigned for each subject group in our experiment. The error bars indicate a 95% confidence interval.

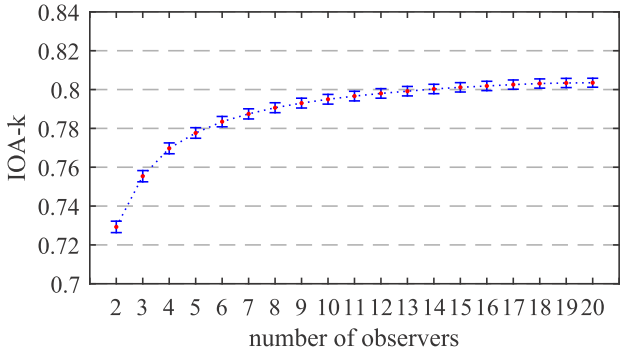


Fig. 5. Illustration of inter-k-observer agreement (IOA-k) value averaged over all stimuli contained in our entire dataset. The error bars indicate a 95% confidence interval.

depending on the choices made in the experimental methodology, such as the sample size and the ways of presenting stimuli [56]. To make use of eye-tracking data as a solid "ground truth", it is crucial to validate the reliability of the collected data. We, therefore, propose and perform systematic reliability testing to assess: (1) whether the variances in the eye-tracking data obtained from different subject groups (in a between-subjects method) are similar; (2) whether the sample size (number of participants) per stimulus is sufficient to create a stable gaze map; and (3) whether the eye-tracking data collected in our study are comparable to similar data obtained from other independent studies. Note, hereafter, when performing a statistical significance test, if the assumption of normality is tested to be satisfied a parametric test (e.g., t-test) is used; otherwise a nonparametric alternative (e.g., Wilcoxon signed rank test) is used.

1) Homogeneity of Variances Between Groups: Since a between-subjects method is adopted, assuming the representativeness of participants in each group is satisfied, we test whether variances of eye-tracking data across all groups are homogeneous. To identify such homogeneity, we measure the inter-observer agreement (IOA), which refers to the degree of agreement in saliency among observers viewing the same stimulus [57], [58]. In our implementation, per stimulus and per subject group, IOA is quantified by comparing the gaze map generated from the fixations over all-except-one observers to the gaze map built upon on the fixations of the excluded observer; and by repeating this operation so that each observer serves as the excluded subject once. The similarity between two gaze maps is commonly measured by AUC (i.e., area under the receiver operating characteristic curve) [13]. Fig. 4 illustrates the IOA value averaged over all stimuli assigned to each subject group in our experiment. It shows that the IOA remains similar across eight groups. A statistical significance test (i.e., analysis of variance (ANOVA)) is performed and the results show that there is no statistically significant difference between groups (i.e., with $P > 0.05$ at the 95% confidence level). The above evaluation indicates that a high degree of consistency across groups is found in our data collection.

2) Data (Saliency) Saturation: There is, unfortunately, no general agreement on how many participants are adequate to achieve reliable eye-tracking data. Researchers often use "data saturation" as a guiding principle to check whether a given/chosen sample size is sufficient to cause a "saturated" gaze map. This means a gaze map reaches the point at which no new information is observed. We test the adequacy of sample size required to reach saliency "saturation" (i.e., a proxy of sufficient degree of reliability) in our experimental data. The validation is again based on the principal of IOA, which is extended to an inter-*k*-observer agreement measure (i.e., referred to as IOA-*k*, and k=2, 3...20). More specifically, for a given stimulus, IOA-*k* is calculated by randomly selecting k participants among all observers. Fig. 5 illustrates the IOA-*k* value averaged over all stimuli contained in our entire dataset. It shows that "saturation" occurs with 16 participants, although a reasonably high degree of consistency in fixation deployment is already reached with 12 participants. It demonstrates that our chosen number of 20 observers for each subject group is fairly sufficient to yield a stable/saturated gaze map.

3) Cross-Database Similarity: To further evaluate the reliability of our eye-tracking data as a "ground truth", we compare our data to other relevant databases that are publicly available and obtained from independent laboratories. In terms of free-viewing eye movement recordings related to the LIVE database, there exist three widely cited eye-tracking databases (with stimuli being only the 29 source images of the LIVE database), namely TUD [29], UN [59] and UWS [30]. An exhaustive comparative study is already conducted in [59], and shows a high degree of similarity between these databases, despite the fact that they were independently collected under different experimental conditions. As a reference provided in [59], for the same image, when comparing its two independently generated gaze maps by means of Pearson correlation, the result that falls into the range [0.8, 0.9] indicates a high degree of similarity. Since we only selected 18 source images from the LIVE database, the comparison had to be based on these 18 images only. The Pearson correlation averaged over all images between our data and TUD is 0.87; and is 0.87 and 0.89 with respect to UN and UWS, respectively. This suggests that our eye-tracking data should be considered as reliable "ground truth".

C. Validation: Impact of Stimulus Repetition

We hereby investigate the impact of stimulus repetition on the reliability of data collection, via a dedicated eye-tracking experiment combining the ideas of both [44] and [45] as

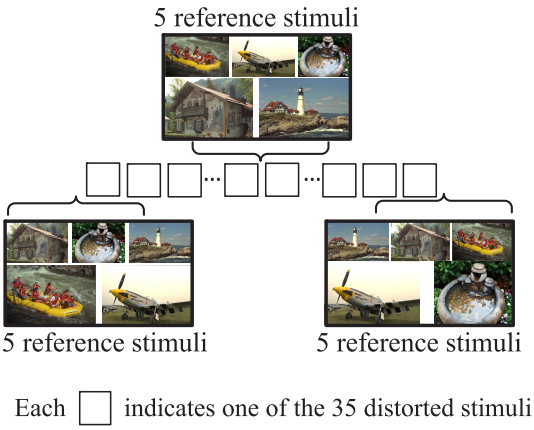


Fig. 6. The construction of stimuli in a single trail. The boxes indicate 35 stimuli in random order. The 5 original images, as a group, are inserted in the front end, middle and back end of each trail in random order.

mentioned in Section II-A. Note our main purpose here is to raise awareness of the need for eliminating stimulus repetition in the scenario where subjects have to view the same scene repeatedly, e.g., 16 times, rather than compare the general usage of different subjective testing methodologies. Our experiment aims to investigate two aspects: 1) how stimulus repetition affects fixation behaviour when viewing several distorted versions of the same scene (as also similarly studied for videos in [45]); 2) how stimulus repetition affects fixation behaviour when viewing several times the same undistorted scene (as also similarly studied for videos in [44]).

We chose five source images to construct our test stimuli. In creating distorted stimuli, we selected 7 distorted images (covering all available distortion types and the full range of DMOS) per content from the LIVE database, resulting in 35 distorted images. In creating undistorted stimuli, we just used the 5 source images three times. This gave a total of 50 test stimuli. As illustrated in Fig. 6, the 35 distorted stimuli were presented in a random order to each participant. The three groups of the same source images (presented in a random order within group) were positioned in the beginning, middle and end of the presentation. Therefore, in terms of the distorted stimuli, there are 7 repetitions per content; and in terms of the undistorted stimuli, there are 3 repetitions per content. We recruited 20 participants (10 females and 10 males) in our experiment. Each participant viewed freely all stimuli. Each stimulus was shown for 10 seconds followed by a mid-grey screen for 3 seconds. We followed the same experimental set-up as described in Section III-C.

1) The Effects for Distorted Stimuli (7 Repetitions): For each participant, first the similarity in fixations between each distorted image and the corresponding source image (presented in the beginning) is measured by AUC. Then, the 7 AUC values per content are ranked in the order of viewing, averaged over all contents and all participants as shown in Fig. 7. It clearly shows the general trend that the similarity decreases as the viewing order increases, independent of the image content, distortion type and distortion level. The results of t-test show that there is a statistically significant difference

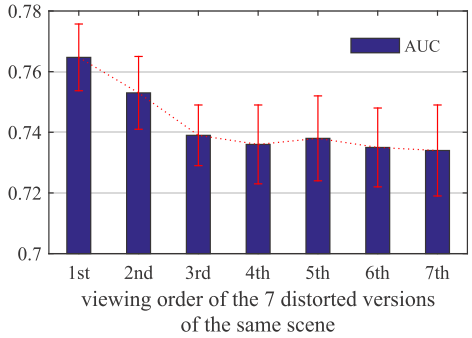


Fig. 7. Illustration of the impact of stimulus repetition on fixation behaviour. When viewing 7 distorted versions of the same scene, the similarity in fixations (measured by AUC) relative to its original decreases as the viewing order increases. The error bars indicate a 95% confidence interval.

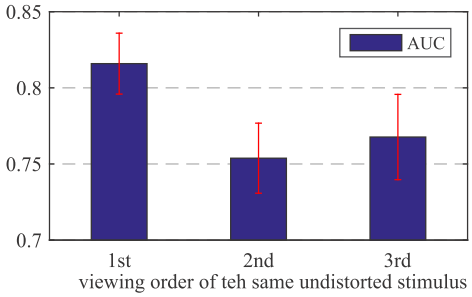


Fig. 8. Illustration of the impact of stimulus repletion on fixation behaviour. When viewing 3 times the same undistorted scene, the similarity in fixations (measured by AUC) relative to its baseline taken from the TUD database decreases as the viewing order increases. The error bars indicate a 95% confidence interval.

between the 1st viewing and the Nth viewing ($N = 3$ to 7) with $P<0.05$ at the 95% confidence level. This suggests that stimulus repetition can significantly impact the fixation behaviour, and consequently bias the intended fixation data.

2) The Effects for Undistorted Stimuli (3 Repetitions): A mean gaze map (over all subjects) is produced for each undistorted stimulus, and is compared by AUC to the corresponding baseline gaze map taken from the TUD database [29]. The gaze maps contained in the TUD database were collected under task-free, no distortion, no stimulus repetition conditions, using the source images of the LIVE database. Fig. 8 illustrates the AUC values in viewing order, averaged over all 5 source images. It shows that the similarity dramatically drops after the first viewing of a scene, independent of image content. A Wilcoxon signed rank test shows that there is a statistically significant difference between the first and the second (or the third) viewing with $P<0.05$ at the 95% confidence level.

The above study provides evidence that when subjects view the same stimuli repeatedly the fixation data are likely to be biased, and care should be taken to eliminate the effect of stimulus repetition in such a scenario.

D. Fixation Deployment

Fig. 9(a) illustrates an overview of all distorted versions (5 distortion types $\times$ 3 distortion levels) of a source image

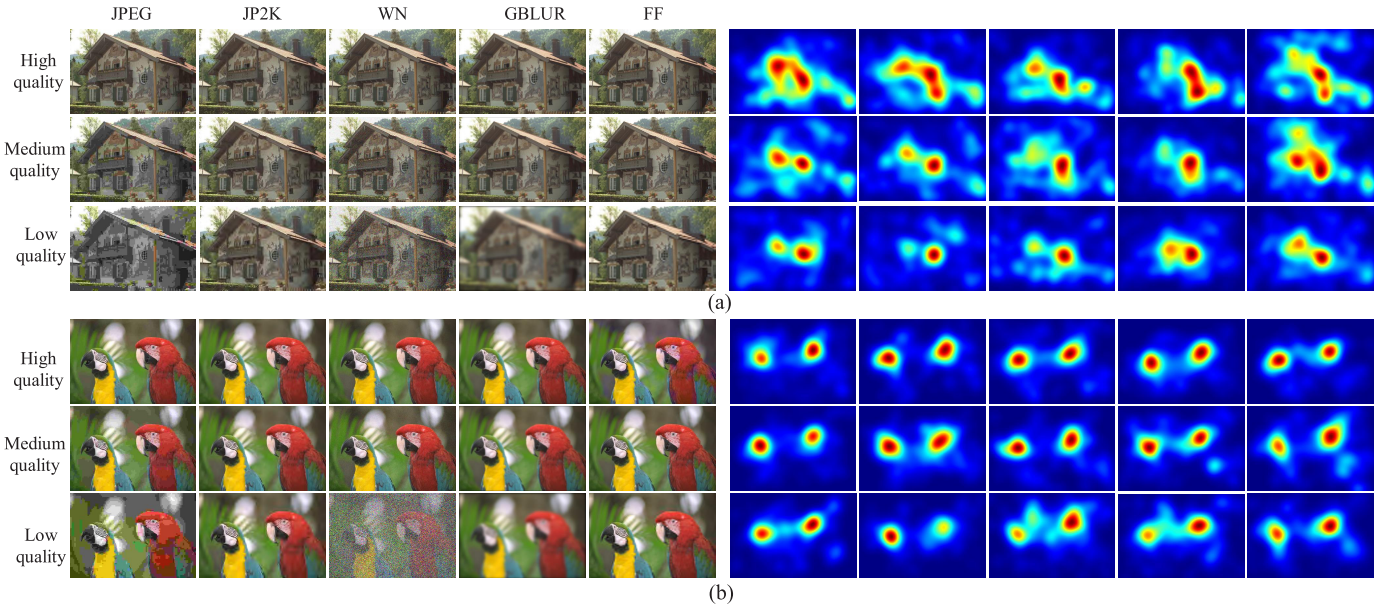


Fig. 9. (a) Illustration of all distorted versions of a source image (of a large degree of saliency dispersion) and their corresponding gaze maps. The same layout of distorted images and gaze maps for a different source image (of a small degree of saliency dispersion) is illustrated in (b).

(of a large degree of saliency dispersion) and their corresponding gaze maps (i.e., referred to as distorted scene saliency (DSS)). The same layout of distorted images and DSS for a different source image (of a small degree of saliency dispersion) is illustrated in Fig. 9(b). The grids visualise typical correspondences and differences between DSS rooted from the same source image. In general, there exist consistent patterns among the relevant DSS, e.g., the highly salient regions tend to cluster around the same positions. However, there are some deviations, which are seemingly caused by either the distortion type or distortion level. It is observed in Fig. 9(a) that as the quality degrades (i.e., the strength of distortion increases) the saliency patterns become more convergent (i.e., less amount of heated areas in DSS); and that at the same distortion level how saliency disperses tends to depend on the distortion type, e.g., at "High" quality saliency is more spread out for JPEG, JP2K and FF than for WN and GBLUR. In addition, the two examples (rooted from two different source images) exhibit different trends in terms of the variation in the array of DSS. For example, the change in quality seems to cause a more obvious rate of convergence in saliency in Fig. 9(a) than in Fig. 9(b). This may be due to the fact that the two source images fall into distinct categories of visual content in terms of saliency dispersion (see Fig. 1). It implies that image content also has an impact on the deployment of DSS, as already mentioned in [31].

V. Interactive Relations Between Saliency and Image Quality Aspects

The resulting eye-tracking data represent sufficient statistical power, which allows further statistical analysis on the observed tendencies in the changes of saliency induced by the changes of image quality aspects. More specifically, we evaluate the impact of three individual categorical variables (i.e., distortion type, distortion level and image content) on the deployment of fixation.

A. Investigation Framework

We use saliency derived from the original undistorted scene (i.e., referred to as scene saliency (SS)) as the reference, and quantify the deviation of DSS from its corresponding reference SS. The deviation between two gaze maps is often quantified by three similarity measures widely used in the literature. They are Pearson linear correlation coefficient (CC) [60], [61], normalized scanpath saliency (NSS) [62], [63] and AUC [13]. The use of these measures is already described in more detail in [64], and we only briefly repeat their meaning in our context as follows:

CC: when CC is close to -1 or 1, the similarity between SS and DSS is high; when CC is close to 0, the similarity is low.

NSS: When NSS$>$0, the higher the value of the measure the more similar DSS and SS are; whereas NSS$<$0 indicates that being able to use DSS to reproduce its reference SS is likely due to chance only.

AUC: AUC$=$1 means DSS can predict perfectly the characteristics of its reference SS; whereas AUC$=$0.5 corresponds to a prediction at chance level.

B. Investigation Results

The statistical evaluation is based on 270 data points (i.e., 270 distorted stimuli rooted from 18 originals) of SS-DSS similarity (i.e., the similarity calculated by CC, NSS and AUC between a given DSS and its corresponding SS). A full factorial ANOVA is conducted with the SS-DSS similarity as the dependent variable (the test for the assumption of normality indicates that the dependent variable is normally distributed); and the distortion type, distortion level and image

TABLE I

RESULTS OF THE ANOVA TO EVALUATE THE IMPACT OF DISTORTION TYPE, DISTORTION LEVEL AND IMAGE CONTENT ON THE MEASURED SIMILARITY BETWEEN SS AND DSS. df DENOTES DEGREE OF FREEDOM, F DENOTES F-RATIO AND Sig DENOTES THE SIGNIFICANCE LEVEL

ANOVA		CC		NSS		AUC	
Source	df	F	Sig	F	Sig	F	Sig
Distortion type	4	2.89	**.02**	1.48	.21	0.92	.45
Distortion level	2	46.7	**.00**	23.44	**.00**	27.89	**.00**
Image content	2	124.33	**.00**	439.96	**.00**	483.7	**.00**
Distortion type * Distortion level	8	2.03	**.04**	1.15	.33	0.95	.48
Distortion type * Image content	8	1.92	**.05**	0.74	.66	0.96	.47
Distortion level * Image content	4	2.82	**.03**	0.1	.98	1.02	.39
Distortion type * Distortion level * Image content	16	0.71	.79	0.32	.99	0.4	.98

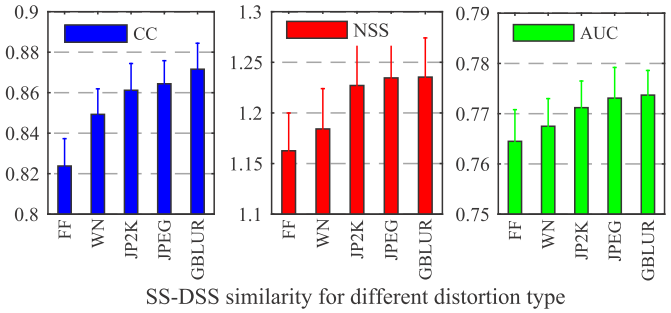


Fig. 10. Illustration of rankings of five distortion types contained in our database in terms of the SS-DSS similarity measured by CC, NSS and AUC, respectively. The error bars indicate a 95% confidence interval.

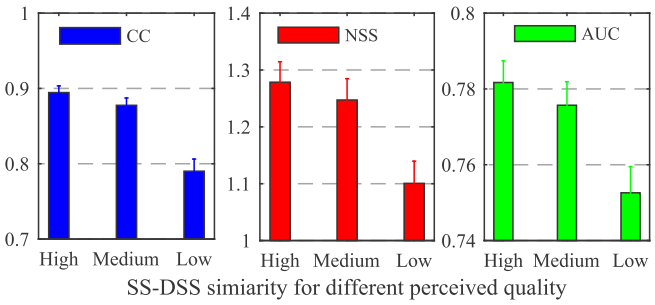


Fig. 11. The measured SS-DSS similarity in terms of CC, NSS and AUC for images of different perceived quality. The error bars indicate a 95% confidence interval.

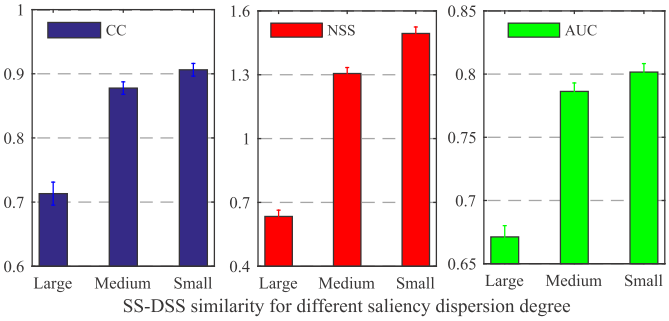


Fig. 12. The measured SS-DSS similarity in terms of CC, NSS and AUC for images of different visual content (i.e., classified by the degree of saliency dispersion). The error bars indicate a 95% confidence interval.

content as independent variables. The results are summarized in Table. I, and show that all main effects (except for the case of distortion type when AUC and NSS are used for SS-DSS similarity) are statistically significant.

1) Impact of Distortion Type on SS-DSS Similarity: As shown in Table I, "distortion type" has a statistically significant effect on SS-DSS similarity measured by CC. The same effect, however, is not found when the SS-DSS similarity is calculated based on NSS or AUC. The inconsistency in the results is attributed to the fact that different similarity measures capture different characteristics of saliency changes while being coherent in measuring SS-DSS similarity, as already mentioned in [64]. CC focuses on the similarity in terms of the spatial distribution of fixation, whereas NSS and AUC are based on the estimation of similarity in terms of the locality and density of fixations. Fig. 10 illustrates the rankings of the five available distortion types in terms of the SS-DSS similarity measured by CC, NSS and AUC, respectively. They consistently produce the same rank order for the five distortion types. For each subplot, the results of hypothesis testing (i.e., Wilcoxon signed rank test) show that the impact of distinct distortion types (e.g., FF and GBLUR) on SS-DSS similarity is statistically different with $P<0.05$ at the 95% confidence level. The distortions contained in FF (i.e., high-frequency, localised artifacts) produce a large extent of saliency deviation, whereas the GBLUR distortions (i.e., low-contrast, uniformly distributed artifacts) cause only slight changes in saliency.

2) Impact of Distortion Level on SS-DSS Similarity: Table I shows that "distortion level" has a statistically significant effect on SS-DSS similarity, independent of the similarity measure used. The degree of saliency deviation increases as the perceived quality decreases (or strength of distortion increases). Fig. 11 illustrates the measured SS-DSS similarity (again in terms of CC, NSS and AUC) for three levels of perceived quality. It reveals a statistically significant (i.e., based on t-test with $P<0.05$ at the 95% confidence level) drop in SS-DSS similarity at low quality relatively to teh other two cases, which means that the distraction power of the annoying artifacts (or strong distortions) present in an image comes into impact the perception of the natural scene.

3) Impact of Image Content on SS-DSS Similarity: Table I also shows that SS-DSS similarity is strongly affected by "image content" (i.e., classified by the degree of saliency dispersion). Fig. 12 illustrates the measured SS-DSS similarity (again in terms of CC, NSS and AUC) for images having different degrees of saliency dispersion. In the case of images

TABLE II

PERFORMANCE FOR 10 OQMs (CC WITHOUT NON-LINEAR FITTING) AND THEIR CORRESPONDING SALIENCY-BASED VERSIONS ON OUR DATABASE WITH 270 DISTORTED STIMULI

	PSNR	UQI	SSIM	MS-SSIM	VIF	FSIM	GBIM	NPBM	JNBM	NBAM	average ΔCC
original	0.784	0.891	0.773	0.791	0.917	0.855	0.809	0.842	0.854	0.828	-
SS-based	0.800	0.910	0.817	0.824	0.922	0.857	0.834	0.862	0.871	0.871	0.022
DSS-based	0.801	0.912	0.818	0.824	0.920	0.858	0.867	0.849	0.855	0.866	0.023

TABLE III

RESULTS OF STATISTICAL SIGNIFICANCE TESTING FOR INDIVIDUAL OQMs. "1" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT WITH $P<0.05$ AT THE 95% CONFIDENCE LEVEL. "0" MEANS THAT THE DIFFERENCE IS NOT SIGNIFICANT

	PSNR	UQI	SSIM	MS-SSIM	VIF	FSIM	GBIM	NPBM	JNBM	NBAM
original vs. SS-based	1	1	1	0	1	0	1	1	1	1
original vs. DSS-based	1	1	1	1	1	1	1	1	1	1

that do not contain highly salient objects (i.e., a large degree of saliency dispersion), adding artifacts to these images results in substantial changes between SS and DSS, as indicated by the statistically significant (i.e., based on t-test with $P<0.05$ at the 95% confidence level) drop in SS-DSS similarity relatively to the other two cases. On the other hand, images with highly salient objects (i.e., a small degree of saliency dispersion) are less sensitive to the distortions, as evidenced by the statistically significantly larger (i.e., based on t-test with $P<0.05$ at the 95% confidence level) values of CC, NSS and AUC.

VI. SS VERSUS DSS ON THE PERFORMANCE GAIN OF OQMs

Previous research [29] has demonstrated that adding "ground truth" SS does improve the performance of OQMs in predicting perceived image quality. The findings, however, also showed that the performance gain could be potentially optimised by taking into account the interactions between SS and distortion. DSS, to some extent, represents the interactive effect of the concurrence of natural scene and unnatural artifacts. The added value of DSS as opposed to SS in OQMs, however, has not been investigated. To provide insights into this matter, both types of saliency are added to several OQMs well-known in the literature.

A. Investigation Framework

We follow the general framework established in [65] for assessing the added value of saliency in OQMs. The basic idea is to quantify the performance gain of an OQM by comparing its predictive power with and without saliency. The predictive power of an OQM can be simply measured by the Pearson correlation (i.e., CC) between the output of the OQM and the subjective quality ratings [66]; and the performance gain can be effectively expressed by the increase in CC (i.e., ΔCC). The OQMs used in our evaluation are six full-reference (FR) OQMs, peak signal-to-noise ratio (PSNR) [1], universal quality index (UQI) [67], structural similarity index (SSIM) [12], multi-scale SSIM (MS-SSIM) [68], visual information fidelity (VIF) [69] and feature similarity index (FSIM) [70]; and four no-reference (NR) OQMs, generalized block-edge impairment metric (GBIM) [71], NR blocking artifact measure (NBAM) [72], NR perceptual blur metric (NPBM) [73] and just noticeable blur metric (JNBM) [74].

B. Investigation Results

1) Original Versus Saliency-Based OQMs: Per OQM, adding SS and DSS (i.e., as the implementation detailed in [65]) results in two new saliency-based OQMs. The performance (i.e., CC) of an OQM is calculated based on the subjective quality scores contained in our database, which is summarised in Table II. In general, it shows that the performance of OQMs is improved by using both SS and DSS. The gain (i.e., ΔCC) ranges from 0.002 (FSIM extended with SS) to 0.058 (GBIM extended with DSS). Note VIF and FSIM obtain relatively small gain by adding saliency, due to the fact that some well-established saliency aspects (i.e., information content feature in VIF [75] and phase congruency feature in FSIM [76]) are already embedded in these metrics, which consequently causes a saturation effect in saliency optimisation [65].

The observed effects are statistically analysed with hypothesis testing, selecting the metric strategy (SS-based vs. original or DSS-based vs. original) as the independent variable and the performance gain as the dependent variable. A Wilcoxon signed rank test is performed using the data points contained in Table II. The results, with $P<0.01$ at the 95% confidence level reveal that both SS and DSS statistically significantly improve the original OQMs. To further check the effectiveness of adding saliency for individual OQMs, the differences were statistically analysed per OQM (i.e., as the implementation detailed in [65]): in the case of normality, t-test was performed; otherwise a Wilcoxon signed rank test was conducted, as the results summarised in Table III.

2) SS-Based Versus DSS-Based OQMs: As can be seen in Table II, on average (over all OQMs), the gain achieved by use of SS is similar to that of using DSS. To check the effects with a statistical analysis, a Wilcoxon signed rank test is performed, selecting the type of saliency as the independent variable and the performance as the dependent variable. The test results (i.e., $p>0.05$ at the 95% confidence level) show that there is no statistically significant difference between the inclusion of both types of saliency.

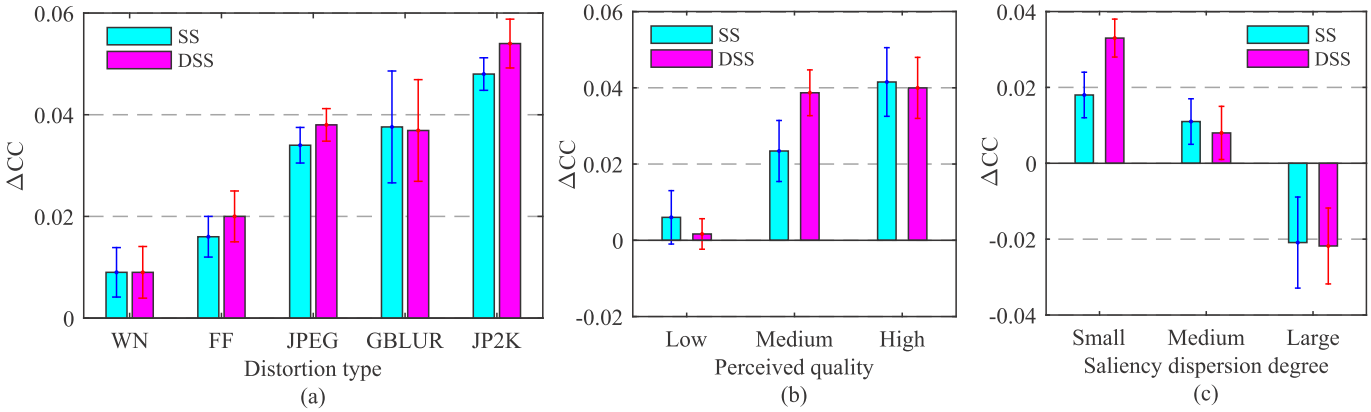


Fig. 13. Comparison of performance gain between SS-based and DSS-based OQMs, with the effect of (a) distortion type dependency, (b) perceived quality level dependency and (c) saliency dispersion degree dependency. The error bars indicate a 95% confidence interval.

In response to the investigation framework identified in Section V, we further assess how the performance gain between SS-based and DSS-based OQMs is affected by the observed main effects, i.e., the distortion type, distortion level and image content. More specifically, our database is again characterised at three individual aggregation levels, using "distortion type", "distortion level" and "image content" as the classification variables, respectively.

Fig. 13(a) illustrates the performance gain (i.e., ΔCC) averaged once over all SS-based OQMs and once over all DSS-based OQMs, when assessing WN, JPEG, GBLUR, JP2K and FF, respectively. It shows that both types of saliency are beneficial for OQMs (i.e., ΔCC values are positive in all cases). Results of a Wilcoxon signed rank test show that the difference in performance gain between the use of SS and DSS is not statistically significant different with $P>0.05$ at the 95% confidence level for all distortion types except for JP2K. For JP2K, using DSS improves the OQMs' performance more, which is in line with the conclusions drawn in [65] that when saliency is added in OQMs for accessing localised distortion, such as JP2K, taking into account the interactions between saliency and distortion can be used to optimise the performance gain. Note the same trend is also observed for the localised JPEG and FF distortion, although the results are not significant in our current samples.

Fig. 13(b) shows the comparison of ΔCC between SS-based and DSS-based OQMs, when accessing images with three distinct levels of perceived quality. At low quality, OQMs do not benefit from the use of saliency (i.e., marginal values of ΔCC). At high quality, there is no statistically significant difference (i.e., based on t-test with $P>0.05$ at the 95% confidence level) between the added value of SS and DSS, which is attributed to the fact that SS and DSS is very similar (i.e., a small degree of SS-DSS deviation as shown in Fig. 11). In terms of the medium level of quality, the results of a t-test (with $P<0.05$ at the 95% confidence level) demonstrate that adding DSS to OQMs yields statistically significantly higher performance gain than adding SS, suggesting that the use of saliency in OQMs potentially benefits from taking into account the interactions between saliency and distortion.

Fig. 13(c) illustrates the difference in ΔCC between SS-based and DSS-based OQMs, when accessing images with three distinct degrees of saliency dispersion. Adding saliency deteriorates the performance of OQMs for assessing images with a large degree of saliency dispersion, which should be avoided in saliency optimisation. This is mainly due to the uncertainty of a dispersed gaze map, which confuses the workings of OQMs by e.g., unhelpfully downplaying the importance of high distortion in certain regions [31]. Images with a medium range of saliency dispersion do not profit from adding saliency to an OQM (i.e., marginal ΔCC). For images having a small degree of saliency dispersion, the use of DSS produces statistically significantly (i.e., based on t-test with $P<0.05$ at the 95% confidence level) larger ΔCC than that of using SS. Again, this suggests the interactions between saliency and distortion play a significant role in optimising the increase in the performance of OQMs.

VII. Study of Modelled SS and DSS

A realistic OQM, however, will use a computational saliency model rather than eye-tracking. Before the application of a saliency model, it is highly desirable to validate its performance against the ground truth. Benchmarking saliency models against ground truth SS has been attempted [13], [65], [77]; however, little is known about the performance of existing saliency models in detecting DSS. Questions still remain whether these saliency models sufficiently cope with the distortions added to the undistorted scenes, or at least whether they operate on the original and the distorted stimuli in a similar manner. In addition, the benefits of including SS versus DSS in OQMs have been demonstrated by use of eye-tracking data in Section VI. It is worthwhile to verify whether the findings still remain significant, and potentially useful, when computational saliency is used in this place.

A. Predictive Power of Saliency Models

Our evaluation is carried out with 27 state of the art saliency models, namely ITTI, Torralba, AIM, STB, GBVS, GR, LC, SR, DVA, SUN, FTS, Judd, SDSR, PQFT, CBS, AWS, Gazit,

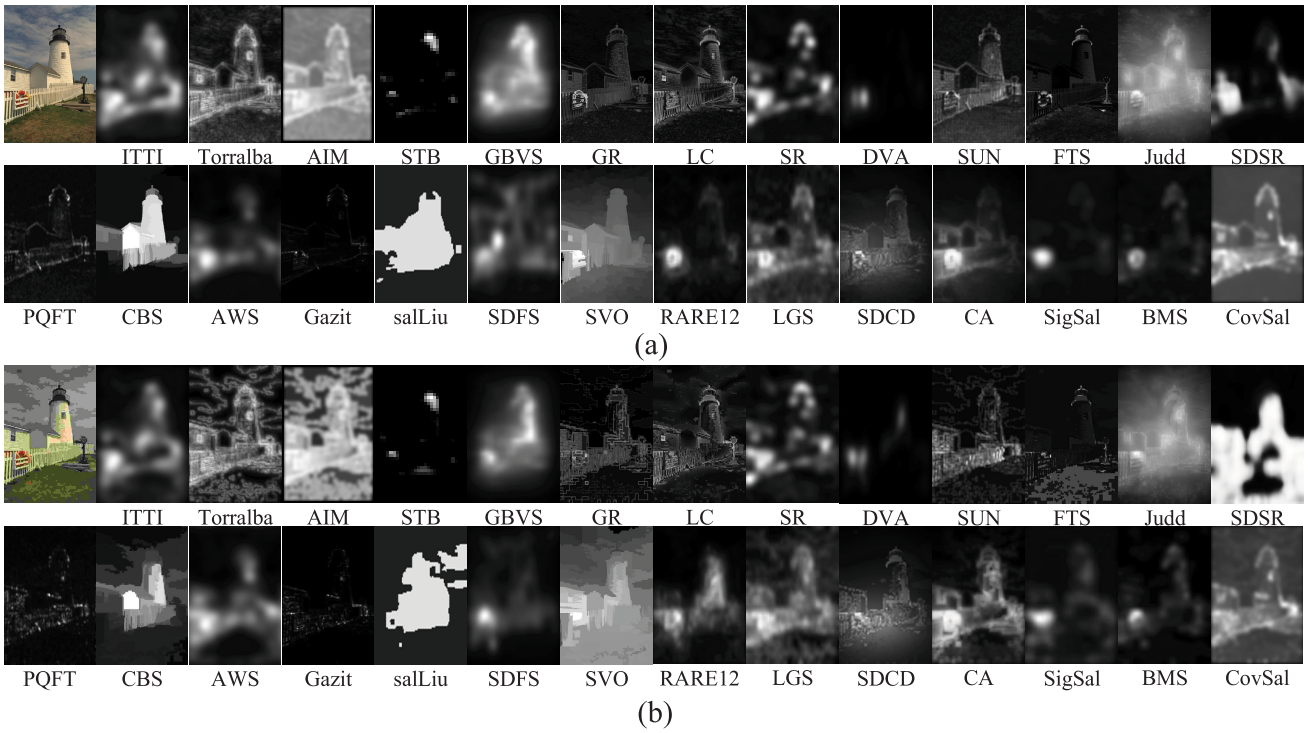


Fig. 14. Illustration of modelled saliency maps generated by twenty-seven saliency models for one of the source images (a) and one of its distorted versions (i.e., JPEG, DMOS=90.43) (b) in our database.

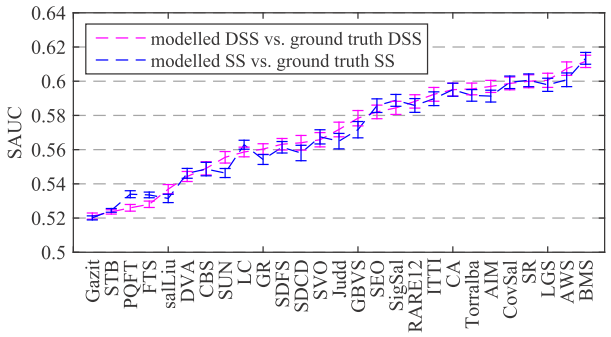


Fig. 15. Illustration of teh predictive power for 27 saliency models. Modelled SS is evaluated against ground truth SS. Modelled DSS is evaluated against ground turth DSS. The error bars indicate a 95% confidence interval.

salLiu, SDFS, SVO, RARE2012, LGS, SDCD, CA, SigSal, BMS and CovSal (as already detailed in [13], [77], and [78]). Fig. 14 shows the modelled saliency maps generated by these models for one of the source images and one of its distorted versions in our database. The predictive power of a saliency model is quantified by SAUC (i.e., shuffled AUC as defined in [13]).

Existing saliency models are usually evaluated against the fixations collected with undistorted natural scene stimuli (i.e., SS), we now check their corresponding performance on distorted stimuli (i.e., modelled DSS). Fig. 15 illustrates the predictive power (i.e., based on SAUC) of the saliency models using the subset of undistorted stimuli and the subset of distorted stimuli in our database, where modelled SS is evaluated against ground truth SS and modelled DSS is evaluated against ground truth DSS. The Pearson correlation between the two sets of SAUC values is 0.98, indicating that the performance of individual saliency models is consistent in both cases. A t-test is also conducted between the two sets of SAUC values; and the results (i.e., with $P > 0.05$ at the 95% confidence level) show that the average performance of saliency models is the same for both cases.

B. Modelled SS Versus DSS in OQMs

We conducted a statistical evaluation using 27 state of the art saliency models and 10 best-known OQMs as used in Section VI. The study thus resulted in 270 saliency-augmented OQMs; and the performance of each OQM was evaluated against the entire LIVE database (with 779 stimuli). In each case, both modelled SS and DSS are generated by applying a saliency model to the reference and distorted image. Table IV shows the performance (i.e. CC without non-linear regression) in each case, averaged over 27 saliency models.

In contrast to the conclusions concerning Table II, Table IV reveals the following consistent findings: (1) OQMs generally benefit from including both modelled SS and DSS. A Wilcoxon signed rank test is performed using the data points of Table IV. The results, with $P<0.05$ at the 95% confidence level, show that both modelled SS and DSS statistically significantly improve the original OQMs. (2) On average (over all OQMs), there is no statistically significant difference between the use of modelled SS and DSS, as demonstrated by a Wilcoxon signed rank test with $P>0.05$ at the 95% confidence level.

We also repeated the same experiment in Section VI to evaluate how the observed effects, i.e., the distortion type, distortion level and image content, impact the optimal use of

TABLE IV

PERFORMANCE OF 10 OQMS (CC WITHOUT NON-LINEAR FITTING) AND THEIR CORRESPONDING SALIENCY-BASED VERSIONS ON LIVE DATABASE WITH 779 DISTORTED STIMULI. NOTE THAT CC IS AVERAGED OVER ALL SALIENCY MODELS

	PSNR	UQI	SSIM	MS-SSIM	VIF	FSIM	GBIM	NPBM	JNBM	NBAM	average ΔCC
original	0.859	0.898	0.825	0.830	0.945	0.859	0.773	0.843	0.833	0.836	-
SS-based	0.870	0.910	0.857	0.859	0.937	0.846	0.803	0.874	0.851	0.857	0.016
DSS-based	0.871	0.906	0.869	0.853	0.942	0.857	0.801	0.869	0.848	0.859	0.017

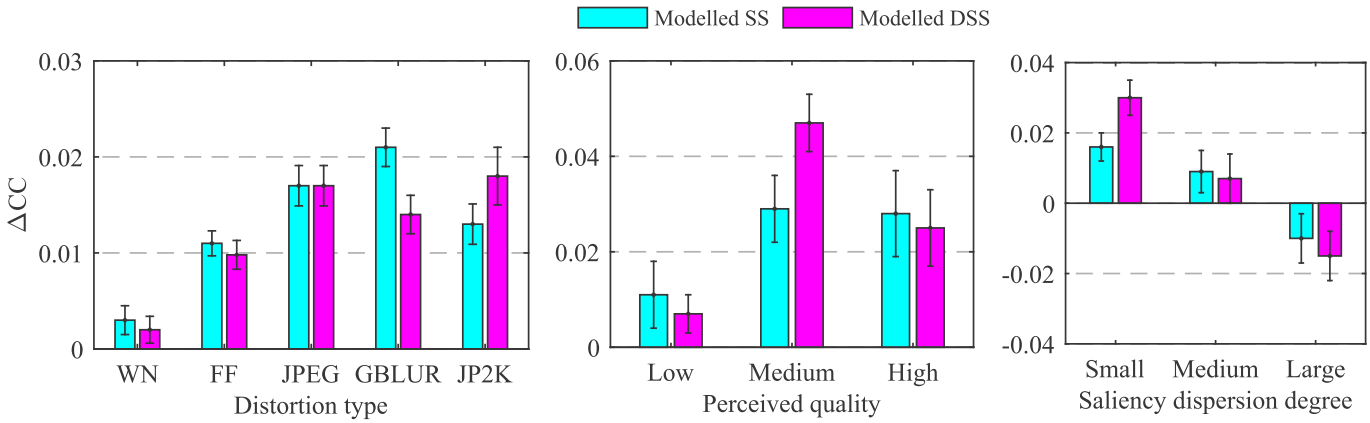


Fig. 16. Comparison of performance gain between modelled SS-based and modelled DSS-based OQMs, with the effect of (a) distortion type dependency, (b) perceived quality level dependency and (c) saliency dispersion degree dependency. The error bars indicate a 95% confidence interval.

modelled SS and DSS. Compared to the results reported in Fig. 13, Fig. 16 shows: (1) In terms of the impact of distortion type, the results of a t-test with $P<0.05$ at the 95% confidence level show that the difference between the use of modelled SS and DSS is not statistically significant for WN, FF, JPEG. For JP2K, modelled DSS yields a statistically significantly (i.e., based on t-test with $P<0.05$ at the 95% confidence level) larger ΔCC than modelled SS. The above findings are consistent with the results determined by eye-tracking data in Fig. 13. For GBLUR, modelled SS produces statistically significant (i.e., based on t-test) larger gain than modelled DSS with $P<0.05$ at the 95% confidence level, which is inconsistent with the results as shown in Fig. 13. The relatively small gain obtained from modelled DSS is mainly caused by the fact that saliency models cannot fully capture the salient features of blurred images (or modelled saliency computed on blurred images is less accurate), which consequently reduces the usefulness of including saliency to an OQM. (2) In terms of the impact of distortion level and image content, Fig. 16 shows the consistent findings as also presented in Fig. 13. We again performed the t-tests on our data. The results show that using modelled DSS in OQMs produces statistically significantly larger gain than using modelled SS, with $P<0.05$ at the 95% confidence level, when assessing the images of medium quality and images having a small degree of saliency dispersion. The observed tendencies can therefore serve as useful tools in optimising the saliency integration in OQMs.

VIII. CONCLUSIONS

In this paper, we investigated a more reliable methodology for collecting eye-tracking data for image quality study. We proposed dedicated control mechanisms to effectively eliminate potential bias due to the involvement of massive stimulus repetition. The refined methodology resulted in a new eye-tracking database with a large degree of stimulus variability, including 288 test images distorted with different types of artifacts at various levels of degradation. The database contains 5760 eye movement trials recorded with 160 human observers.

Based on the "ground truth" data, we thoroughly assessed the interactions between saliency and distortion. An exhaustive statistical evaluation was conducted to provide insights into the tendencies in the changes of saliency induced by distortion. We found that the occurrence of distortion in an image tends to deviate fixation deployment. We also quantified the extent of such deviation as a function of distortion type, degradation level and image content, respectively. In terms of optimal use of saliency in OQMs, we investigated whether saliency of the undistorted scene or that represents the same scene affected by distortion would deliver the best performance gain for OQMs. The results show that both types of saliency are beneficial for OQMs, but the latter which reflects the interactions between saliency and distortion tends to further boost the effectiveness of the integration of saliency in OQMs.

We make use of our new eye-tracking database to benchmark saliency models for the purpose of image quality assessment. The evaluation indicates that existing saliency models operate on the undistorted and distorted scenes in a similar manner in terms of predicting human fixations. Moreover, the findings regarding the benefits of including SS versus DSS in OQMs still hold when using computational saliency instead of eye-tracking data.

Avenues for future research include an in-depth understanding of how visual attention plays a role in assessing image quality, and a quest for a perceptually optimised saliency integration strategy for quality assessment applications.

REFERENCES

[1] Z. Wang and A. C. Bovik, *Modern image quality assessment*. San Rafael, CA, USA: Morgan & Claypool, 2006.

[2] P. C. Cosman, R. M. Gray, and R. A. Olshen, "Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy," *Proc. IEEE*, vol. 82, no. 6, pp. 919–932, Jun. 1994.

[3] Z. Wang, "Applications of objective image quality assessment methods [applications corner]," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 137–142, Nov. 2011.

[4] A. B. Watson, *Digital Images and Human Vision*. Cambridge, MA, USA: MIT Press, 1997.

[5] B. A. Wandell, *Foundations of Vision*. Sunderland, MA, USA: Sinauer Associates, 1995.

[6] W. S. Geisler and M. S. Banks, "Visual performance," in *Handbook of Optics*. New York, NY, USA: McGraw-Hill, 1995.

[7] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1164–1175, Aug. 1997.

[8] S. J. Daly, "Visible differences predictor: An algorithm for the assessment of image fidelity," *Proc. SPIE*, vol. 1666, pp. 2–15, Aug. 1992.

[9] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA, USA: MIT Press, 1993.

[10] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," in *24th Soc. Inf. Dispaly Dig. Tech. Papers*, 1993, p. 946.

[11] J. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 525–536, Jul. 1974.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[13] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, Jan. 2013.

[14] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2049–2056.

[15] A. A. Salah, E. Alpaydin, and L. Akarun, "A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 420–425, Mar. 2002.

[16] C. Breazeal, "A context-dependent attention system for a social robot," in *Proc. 16th Int. Joint Conf. Artif. Intell.*, 1999, pp. 1146–1153.

[17] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[18] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.

[19] D. T. Levin and D. J. Simons, "Failure to detect changes to attended objects in motion pictures," *Psychonomic Bull. Rev.*, vol. 4, no. 4, pp. 501–506, Dec. 1997.

[20] J. E. Raymond, K. L. Shapiro, and K. M. Arnell, "Temporary suppression of visual processing in an rsvp task: An attentional blink?" *J. Experim. Psychol., Human Perception Perform.*, vol. 18, no. 3, pp. 849–860, 1992.

[21] S. Treue, "Neural correlates of attention in primate visual cortex," *Trends Neurosci.*, vol. 24, no. 5, pp. 295–300, 2001.

[22] E. T. Rolls and G. Deco, "Attention in natural scenes: Neurophysiological and computational bases," *Neural Netw.*, vol. 19, no. 9, pp. 1383–1394, 2006.

[23] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[24] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, nos. 1–2, pp. 507–545, Oct. 1995.

[25] B. Fischer and B. Breitmeyer, "Mechanisms of visual attention revealed by saccadic eye movements," *Neuropsychologia*, vol. 25, no. 1, pp. 73–83, 1987.

[26] J. M. Henderson, "Visual attention and eye movement control during reading and picture viewing," in *Eye Movements and Visual Cognition*. New York, NY, USA: Springer, 1992, pp. 260–283.

[27] J. E. Hoffman and B. Subramaniam, "The role of visual attention in saccadic eye movements," *Perception Psychophys.*, vol. 57, no. 6, pp. 787–795, 1995.

[28] K. Rayner, "Eye movements and attention in reading, scene perception, and visual search," *Quart. J. Experim. Psychol.*, vol. 62, no. 8, pp. 1457–1506, 2009.

[29] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 971–982, Jul. 2011.

[30] U. Engelke, H. Kaprykowsky, H. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 50–59, Nov. 2011.

[31] H. Liu, U. Engelke, J. Wang, P. Le Callet, and I. Heynderickx, "How does image content affect the added value of visual attention in objective image quality assessment?" *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 355–358, Apr. 2013.

[32] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency based objective quality assessment of decoded video affected by packet losses," in *Proc. 15th IEEE Int. Conf. Image Process.*, San Diego, CA, USA, Oct. 2008, pp. 2560–2563.

[33] N. G. Sadaka, L. J. Karam, R. Ferzli, and G. P. Abousleman, "A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling," in *Proc. 15th IEEE Int. Conf. Image Process.*, San Diego, CA, USA, Oct. 2008, pp. 369–372.

[34] Q. Ma and L. Zhang, "Image quality assessment with visual attention," in *Proc. 15th Int. Conf. Pattern Recognit.*, Tampa, FL, USA, Dec. 2008, pp. 1–4.

[35] R. Barland and A. Saadane, "Blind quality metric using a perceptual importance map for JPEG-20000 compressed images," in *Proc. 13th IEEE Int. Conf. Image Process.*, Atlanta, GA, USA, Oct. 2006, pp. 2941–2944.

[36] D. Venkata Rao, N. Sudhakar, I. R. Babu, and L. P. Reddy, "Image quality assessment complemented with visual regions of interest," in *Proc. Int. Conf. Comput., Theory Appl.*, Mar. 2007, pp. 681–687.

[37] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 2. Sep. 2007, pp. II-169–II-172.

[38] E. C. Larson and D. M. Chandler, "Unveiling relationships between regions of interest and image fidelity metrics," *Proc. SPIE*, vol. 6822, pp. 68222A-1–68222A-16, Jan. 2008.

[39] E. C. Larson, C. Vu, and D. M. Chandler, "Can visual fixation patterns improve image fidelity assessment?" in *Proc. 15th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2008, pp. 2572–2575.

[40] C. T. Vu, E. C. Larson, and D. M. Chandler, "Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation*, Mar. 2008, pp. 73–76.

[41] X. Min, G. Zhai, Z. Gao, and C. Hu, "Influence of compression artifacts on visual attention," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2014, pp. 1–6.

[42] J. Redi, H. Liu, R. Zunino, and I. Heynderickx, "Interactions of visual attention and quality perception," *Proc. SPIE*, vol. 7865, pp. 78650S-1–78650S-11, Feb. 2011.

[43] U. Engelke, H.-J. Zepernick, and A. Maeder, "Visual fixation patterns in subjective quality assessment: The relative impact of image content and structural distortions," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst.*, Dec. 2010, pp. 1–4.

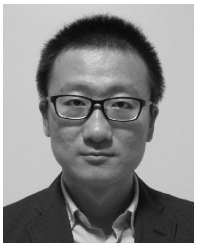
[44] H. Hadizadeh, M. J. Enriquez, and I. V. Bajic, "Eye-tracking database for a set of standard video sequences," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 898–903, Feb. 2012.

[45] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric," *Signal Process., Image Commun.*, vol. 25, no. 7, pp. 547–558, Aug. 2010.

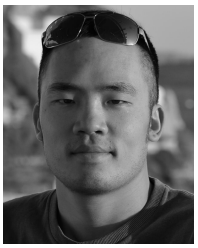
[46] W. Zhang, J. V. Talens-Noguera, and H. Liu, "The quest for the integration of visual saliency models in objective image quality assessment: A distraction power compensated combination strategy," in *Proc. 22nd IEEE Int. Conf. Image Process.*, Quebec City, QC, Canada, Sep. 2015, pp. 1250–1254.

[47] J. Redi, H. Liu, P. Gastaldo, R. Zunino, and I. Heynderickx, "How to apply spatial saliency into objective metrics for JPEG compressed images?" in *Proc. 16th IEEE Int. Conf. Image Process.*, Cairo, Nov. 2009, pp. 961–964.

[48] W. Zhang and H. Liu. (2017). *Toolbox: Integration of Visual Saliency in Objective Image Quality Assessment*. [Online]. Available: https://sites.google.com/site/vaqatoolbox/
[49] H. R. Sheikh, Z. Wang, L. Cormack, and A. Bovik. *LIVE Image Quality Assessment Database Release 2*. [Online]. Available: http://live.ece.utexas.edu/research/quality
[50] D. S. Wooding, "Eye movements of large populations: Ii. deriving regions of interest, coverage, and similarity using fixation maps," *Behavior Res. Methods, Instrum. Comput.*, vol. 34, no. 4, pp. 518–528, Nov. 2002.
[51] M. Mancas and O. Le Meur, "Memorability of natural scenes: The role of attention," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 196–200.
[52] A. G. Greenwald, "Within-subjects designs: To use or not to use?" *Psychol. Bull.*, vol. 83, no. 2, pp. 314–320, 1976.
[53] G. Keren, "Between or within subjects design: A methodological dilemma," *A Handbook for Data Analysis in the Behavioral Sciences*, 2014, pp. 257–272.
[54] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, Standard BT.500-11, International Telecommunication Union, Geneva, Switzerland, 2002, pp. 53-–56.
[55] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Symp. Eye Tracking Res. Appl.*, Tampa, FL, USA, 2000, pp. 71–78.
[56] S. Winkler and R. Subramanian, "Overview of eye tracking datasets," in *Proc. 5th Int. Workshop Quality Multimedia Exp. (QoMEX)*, Klagenfurt, Austria, Jul. 2013, pp. 212–217.
[57] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychol. Rev.*, vol. 113, no. 4, pp. 766–786, 2006.
[58] T. Judd, F. Durand, and A. Torralba, "Fixations on low-resolution images," *J. Vis.*, vol. 11, no. 4, p. 14, 2011.
[59] U. Engelke *et al.*, "Comparative study of fixation density maps," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1121–1133, Mar. 2013.
[60] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.
[61] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "GAFFE: A gaze-attentive fixation finding engine," *IEEE Trans. Image Process.*, vol. 17, no. 4, pp. 564–573, Apr. 2008.
[62] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth, "Variability of eye movements when viewing dynamic natural scenes," *J. Vis.*, vol. 10, no. 10, p. 28, 2010.
[63] Q. Zhao and C. Koch, "Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost," *J. Vis.*, vol. 12, no. 6, p. 22, Jun. 2012.
[64] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1153–1160.
[65] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1266–1278, Jun. 2016.
[66] "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II (FR-TV2)," Video Quality Experts Group, Tech. Rep., 2000.
[67] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[68] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2. Nov. 2003, pp. 1398–1402.
[69] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[70] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[71] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Process. Lett.*, vol. 4, no. 11, pp. 317–320, Nov. 1997.
[72] R. Muijs and I. Kirenko, "A no-reference blocking artifact measure for adaptive video processing," in *Proc. 13th Eur. Signal Process. Conf.*, Antalya, Turkey, Sep. 2005, pp. 1–4.
[73] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. 9th IEEE Int. Conf. Image Process.*, Rochester, NY, USA, Sep. 2002, pp. 57–60.
[74] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 717–728, Apr. 2009.
[75] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.
[76] L. Ma, J. Tian, and W. Yu, "Visual saliency detection in image using ant colony optimisation and local phase coherence," *Electron. Lett.*, vol. 46, no. 15, pp. 1066–1068, Jul. 2010.
[77] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. 12th Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 414–429.
[78] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. (Nov. 2014). "Salient object detection: A survey." [Online]. Available: https://arxiv.org/abs/1411.5878



Wei Zhang (S'14) received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 2011 and 2013, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. His research interests include image analysis, video processing, and human visual perception.



Hantao Liu (S'07–M'11) received the M.Sc. degree from The University of Edinburgh, Edinburgh, U.K., in 2005, and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2011. He is currently an Assistant Professor with the School of Computer Science and informatics, Cardiff University, Cardiff, U.K. His research interests include visual media quality assessment, visual attention modeling and applications, visual scene understanding, and medical image perception. He is currently serving for the IEEE MMTC as the Chair of the Interest Group on Quality of Experience for Multimedia Communications, and an Associate Editor of the IEEE TRANSACTIONS ON HUMAN–MACHINE SYSTEMS.